Secondment report

TIRAMISU DC3.4

Fellow:	Hafsa Tanveer
Research project title:	"Neural Architecture Search framework for efficient and reliable hybrid CNN-Transformer models for Edge AI"
From:	TUT
To:	IFAG
Secondment period:	04.07.2026 – 08.08.2026

Activities during the secondment

Scope and objectives

- Familiarization with the Infineon Edge AI deployment and model optimization workflow.
- Investigation of INT8 post-training quantization (PTQ) and post-hoc reliability calibration for resource-constrained Edge AI models.
- Evaluation of CNN and Vision Transformer models considering three key deployment objectives: classification accuracy, model size, and reliability.
- Evaluation of multiple confidence calibration techniques and calibration-data selection for improving model reliability.

Activities

- Worked with the Infineon model optimization and INT8 PTQ pipeline for efficient Edge AI deployment.
- Evaluated ResNet-8 and ViT-Tiny models on CIFAR-10 using multiple confidence calibration techniques, including Temperature Scaling, Classwise Affine, Matrix Scaling, Structured Matrix Scaling, and Confusion-Guided Residual calibration.
- Performed hyperparameter optimization and candidate analysis using deployed accuracy, INT8 TFLite model size, and reliability as the main evaluation criteria.
- Compared calibration settings with different fitting-data strategies, including full-data, random-subset, and reliability-aware sample selection.

Main results achieved

- Established an end-to-end experimental workflow covering model selection, INT8 PTQ deployment, post-hoc calibration, and reliability evaluation for CNN and ViT models.
- Multiple confidence calibration techniques were successfully implemented and evaluated.
- Quantified model performance using Top-1 accuracy, deployed INT8 TFLite model size, and reliability metrics including NLL, Brier score, ECE, HCE, and a composite reliability risk.
- Established a controlled framework for comparing confidence calibration methods and calibration-data selection strategies.

Next steps

- Validate the observed calibration-data selection behavior on additional CNN and Vision Transformer architectures.
- Extend ARCS with calibration-aware information and compare it with random calibration-data selection using the same number of samples.
- Consolidate the experimental results and prepare a collaborative research publication with Infineon.

Self-evaluation

Overall score

I consider this secondment successful, with regard to the research objectives achieved, skills developed, supervision quality and diversity of the resources. (Agree = 5 ... Disagree = 1)

5

Optional comments

None

Date of the report approval by the supervisor: _____

The TIRAMISU Doctoral Network has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101169378.